Chapter 11 – Sampling Experiments and the Law of Averages

Recall that the box in a *Box Model* represents the Population. In this chapter, we will pretend that we *know the population parameters* in order to try to understand what will then happen in any given sample. In subsequent chapters, we will reverse this process (which is the direction we need in practice).

Example 1. Suppose we know that 63% of Loyola UG students are F, and we take a random sample of size 18 students – what should we expect in our sample? (Notice that this is very artificial since in practice we never know the true $\pi = 63\%$). The box here has one ticket for each UG Loyola student with a 0 (for Males) or 1 (Females) on it, and we take n = 18 tickets (representing the sample) from the Box. Although sampling is done *without replacement* here, since the population is large, one can ignore the SPRF (p.158) and use Binomial calculations anyway.

Note that samples of size 18 will vary in terms of the count and proportion of F's in each sample. These sample proportions (p's) will have the average (or expected value) of $\pi = 0.63$ and Standard Error = $SE_p = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{0.63*0.37}{18}} = 0.1138$. Moreover, the distribution of p's looks approximately Normal.

Example 2. What should we expect to see if we roll a fair die n = 12 times and add the faces? Here, the box contains 6 tickets with the numbers '1' … '6' on them, sampling is done *with replacement*, the sample size (number of draws from the box) is n = 12, and the sample sum is the sample statistic. Each roll of the die (X) has a mean of $\mu_X = 3.5$ and variance of $\sigma_X^2 = 35/12$. Hence, the sum of the 12 rolls (Y) has the average (or expected value) of $\mu_Y = 42$ and the variance of $\sigma_Y^2 = 35$, so SE_{SUM} = SQRT(35) = 5.91608. As above, the sums (Y) have a Normal shape.

Example 3. Suppose that light bulb life follows an Exponential distribution with mean 1 year. We take a sample of size n, and find the sample mean, $\bar{x}$. What should we expect in our sample regarding our sample mean? The average or expected value of the $\bar{x}$'s is 1, the SE is $SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{n}}$, and the shape is approximately Normal. See p.307 in the text and p.3 in the HO (we take n = 4, 9, 16, and 25).

Some Definitions – A *Simple Random Sample* of size n (SRS, p.296) is denoted $X_1 \ldots X_n$, and a (sample) *Statistic* is a random variable calculated from the sample. Whereas *population parameters* (e.g., proportion of F UG students at Loyola) are *numbers*, sample *proportions and other statistics* are *random variables*, and thus vary from sample to sample – that is, they have a *sampling distribution*, which has an average (or expected value) and a standard error. See p.299 top. The difference between the realization of the statistic and the expected value is called the *Random Sampling Error*.

11.3. Law of Averages for Sample Counts – The box contains tickets with only 0's and 1's on them, so that the proportion of 1's (π) corresponds to the probability of a success. The average of the box is π and the SD is $\sqrt{\pi(1-\pi)}$. The n draws from the box corresponds to the sample of size n, and this is done either with or without replacement according to the specific situation; the count (K) is the sum of the tickets drawn.

1. In both cases, the *expected value* of K is $\mu_K = n\pi$;
2. The SE_K is either the SD of the Binomial or Hypergeometric distribution depending on whether sampling is with or without replacement; thus, $SE_K = \sqrt{n\pi(1-\pi)}$ or $\sqrt{\frac{N-n}{N-1}}\sqrt{n\pi(1-\pi)}$;
3. Finally, calculations of probabilities use either exact formulas or Normal or Poisson approximations. When n is large {i.e., both $n\pi$ and $n(1-\pi)$ are at least 5}, the *shape* is approximately Normal.

11.4. Law of Averages for Sample Sums – The box contains tickets each with one value of the possible choices (no longer 0's and 1's). (This is a very theoretical concept. For example, if we measure lifetime lengths of light bulbs, then there is a ticket in the box with *each possible value of bulb length*. If we measure SBP in hepatitis patients, then the box contains tickets with SBP values *for every conceivable hepatitis patient*.) Nonetheless, the mean and SD of the box are μ and σ respectively, and we take a sample of size n from the box – either with or without replacement. Denote the sum in our sample by 'SUM'. Then, we have the following:

1. **The *expected value* of SUM in our sample of size n is $\mu_{SUM} = n\mu$;**
2. **The SE of SUM (SE_{SUM}) is either $\sqrt{n}\sigma$ (for the 'with replacement' case) or $\sqrt{\frac{N-n}{N-1}}\sqrt{n}\sigma$ (for the 'without replacement' case);**
3. **The *shape* is either exactly Normal (when the underlying distribution is Normal) or approximately Normal (for large n).**

Examples: Example 2 above and pp.306-9.

11.5. Law of Averages for Sample Proportions – Since the sample proportion (P) is related to the count (K) by the relation P = K/n, it's easy to show that we have the following:

1. **The *expected value* of P is $\mu_P = \pi$;**
2. **The SE (SE_P) is either $\sqrt{\frac{\pi(1-\pi)}{n}}$ (for the 'with replacement case) or $\sqrt{\frac{N-n}{N-1}}\sqrt{\frac{\pi(1-\pi)}{n}}$ (for the 'without replacement' case);**
3. **Finally, calculations of probabilities use either exact (Binomial or Hypergeometric) formulas or Normal or Poisson approximations. When n is large {i.e., both $n\pi$ and $n(1-\pi)$ are at least 5}, the *shape* is approximately Normal.**

Examples: Example 1 above and pp.313-14.

11.6. Law of Averages for Sample Means – Since the sample mean ($\bar{X}$) is related to the sum ('SUM') by the relation $\bar{X}$ = SUM/n, it's easy to show that we have the following:

1. **The *expected value* of $\bar{X}$ is $\mu_{\bar{X}} = \mu$;**
2. **The SE ($SE_{\bar{X}}$) is either $\frac{\sigma}{\sqrt{n}}$ (for the 'with replacement case) or $\sqrt{\frac{N-n}{N-1}}\frac{\sigma}{\sqrt{n}}$ (for the 'without replacement' case);**
3. **The *shape* is either exactly Normal (when the underlying distribution is Normal) or approximately Normal (for large n).**

Examples: Example 3 above and pp.317-18.

11.7. The Z Statistic – When σ is known, we standardize the *sample mean* by the formula:

$$\mathbf{Z} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

Then, Z has zero mean, SD of 1, and approximately Normal shape.

11.8. The T Statistic – When σ is unknown, we use 's' (the sample SD) in its place and the above formula becomes:

$$\mathbf{T_{n-1}} = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

This statistic has a T distribution with (n-1) *degrees of freedom (df)* – the graph on p.322 shows that the T distribution appears very similar to the Normal distribution, and indeed as n $\rightarrow\infty$ the former approaches the latter. The Z distribution is tabled in the Book front cover and pp. 508-9; the t distribution is tabled in the Book back cover and p.510. The rows in the latter table correspond to the 'df'.

Section 11.9 discusses accuracy, precision, bias, MSE.